Report on Model Assessments for Reading: 2000 - 2001

Prepared by

James C. Impara, Ph.D.

Chad W. Buckendahl, Ph.D.

Barbara S. Plake, Ph.D.

Buros Institute for Assessment Consultation and Outreach

A Division of the

Buros Center for Testing

University of Nebraska - Lincoln

September 2001

Report on Model Assessments for Reading: 2000 - 2001

Organization of this report

This report is organized in three sections. The first section describes briefly the evaluation and review process that identified the model assessments, the format of the model descriptions, and information that will help readers apply the models. The second section identifies districts whose portfolios were used as a base to create the various models. The third section provides a concise description of each model.

Evaluation and review process

Background

The legislation that permits Nebraska school districts to design and use their own unique assessment system for determining the achievement levels of their students on the Nebraska content standards in reading, writing (for some of the standards), listening, and speaking has certain requirements. One such requirement is that the Nebraska Department of Education (NDE) identify four models of assessment that districts may adopt or adapt in designing their future assessments. The NDE employed the Buros Institute for Assessment Consultation and Outreach (BIACO) to assist in the identification of the models. An early decision by the NDE was that the quality of district assessments would be evaluated on the extent that six quality criteria for assessments were met. These quality criteria are: 1) Assessments match the standards; 2) Students have an opportunity to learn assessment content; 3) Assessments are unbiased and free of sensitive language; 4) Assessments are developmentally and cognitively appropriate; and 5) Assessments' scores are consistent/reliable; and 6) Mastery levels are set appropriately.

In discussions between BIACO and NDE it was decided that the legislative requirement for four models could be interpreted to mean four models for each of these six quality criteria. Thus, the BIACO evaluation would ultimately result in 24 model assessment strategies. That is, there would be four models for each of the six quality criteria. A rubric was developed that indicated what might be expected for each quality criterion if the criterion was met.

Sixteen evaluators, named the District Assessment Evaluation Team (DAET), were recruited nationally to apply the rubric to district assessment portfolios. All had some background and training in assessment. Many were either from Nebraska or were familiar with Nebraska school districts. Some had assisted districts in developing their local assessment systems. DAET members who assisted districts in developing district assessments were required to evaluate assessment portfolios of districts they had not assisted. Conflicts of interest were, to the extent possible, eliminated.

The DAET were trained in May 2000 to evaluate District Assessment Portfolios. Several districts that had completed drafts of their portfolios provided them for use in the training. The training occurred over a three-day period.

When districts submitted their final Assessment Portfolio to the NDE, they were resent to BIACO. BIACO repackaged the Assessment Portfolios and sent them to the DAET. Each DAET member reviewed approximately 25 unique Assessment Portfolios. In addition to the

DAET evaluators, BIACO staff also reviewed a large number of Assessment Portfolios. As a quality control measure, two Assessment Portfolios were copied and sent to all reviewers. The evaluations of these showed substantial consistency in the ratings of each of the quality criteria across all DAET and BIACO reviewers.

One element of the DAET review of each district's Assessment Portfolio was to indicate if any of the six quality criteria had been met in an exemplary way. If so, that component of the Assessment Portfolio was identified as a potential model strategy. When all districts and quality criteria were considered, over 75 districts were identified as having potential model strategies for meeting at least one quality criterion. The district Assessment Portfolios that had potential model strategies were set aside for additional review.

Model Selection

The final review to identify model strategies to meet the six quality criteria was accomplished by a National Advisory Committee for Assessment (NACA). In addition to the DAET, a small group of nationally recognized experts in assessment, all of whom have experience or interests in the development and use of local assessments were identified and invited to participate in both the early planning stages and in the selection of models. Two of the NACA members are in a university setting, one is in a state education agency, and the fourth is in a local school district. None have worked with school districts in Nebraska.

In late August, after all DAET reviews were completed, the NACA was convened to review the district assessment portfolios identified by the DAET as having possible model strategies. This review process took about 2 days.

In the process of selecting model strategies, it was noted that the descriptions in some assessment portfolios were incomplete (most often the process was described, but results of the process used were not reported), so the NACA members expanded the descriptions based on the likely results of the process. Thus, the descriptions of the models reflect district procedures in general, but the district(s) listed as being illustrative of that model may not have done everything described for that model strategy.

It is also important to note that over 325[1] unique assessment portfolios were submitted. Each DAET member read between 20 and 25 of the portfolios. It is very possible that a district that was named as being illustrative of a model was not the only district that used the same process as described in the model. In some cases, the NACA noted that several districts (or an identified consortium of districts) used these procedures for a particular quality criterion, but even in these cases, it is likely that other districts that are not named used these same procedures.

One objective of the process of identifying models and districts associated with those models was to identify as many different districts as possible. Thus, there were districts that used exemplary strategies across several of the quality criteria, but are named explicitly only once. One consequence of this is that, for some of the quality criteria it may have been possible to identify more than four models. A second consequence is that a district that did model work on several criteria may be named only once. Some districts did not undertake strategies to meet

[1] Many districts were included in the assessment portfolios submitted by a consortium. There may have been only a single portfolio submitted for the entire consortium. This is why there are fewer assessment portfolios than there are districts.

some criteria. However, they did describe what they planned to do to meet the criteria in their next assessment. These districts were not identified as models.

Within each of the quality criteria, an attempt was made to identify model strategies from districts with different characteristics. This means that, to the extent possible, small, medium, and large districts were identified as models for each of the quality criteria. Moreover, there was an attempt to identify four different strategies for each criterion. In some cases, the variation among the four models was only slight. For example, one district may have used a panel of local teachers to judge the match of the assessment to the standards and another district may have used a panel of local teachers supplemented by teachers from another district (i.e., used resources from outside the district).

The rating scale for each criterion was Met; Met - With Comments; Met - Needs Improvement; and Not Met. Each of the model strategies would be classified as Met, assuming that appropriate description of the procedures and the results of the procedures were provided.

In general, districts named as illustrative of the model strategies provided reasonable descriptions of what they did to meet the criterion and many provided results of their procedures. In many cases, the model strategies for quality criteria 1 through 4 included some element of professional development and often used more than one procedure to meet the standard. Thus, districts used multiple methods to verify that the criterion was being met.

In summary, a rubric was developed to define the quality criteria for assessment of Nebraska's content standards in Reading/Writing. A team of 16 individuals who have strong backgrounds in assessment were trained to rate district assessment portfolios and to identify potential model strategies associated with each of these six quality criteria. A National Advisory Committee came together to review the potential strategies and to provide descriptions of 24 model strategies that districts may consider when revising their assessments in the future.

Districts that were identified as illustrative of the model described in the next section.

Criterion	District(s) that are illustrative of the model	Label
1	Bellevue	1A
1	Beatrice & Omaha	1B
1	Hampton	1C
1	Waterloo	1D
2	Elkhorn	2A
2	Crawford	2B
2	Ralston	2C
2	Hanover & Nemaha Valley	2D
3	Valley, Blair, Bennington, Conestoga, Johnson-Brock	3A
3	Lincoln	3B
3	Winnebago	3C
3	David City	3D
4	Bancroft-Rosalie, Mead, Scribner-Snyder	4A
4	Crete, Ashland-Greenwood	4B
4	Hastings	4C
4	McCool Junction	4D
5	Niobrara	5A
5	Millard	5B
5	Cheney	5C
5	Raymond Central, Yutan	5D
6	Ralston	6A
6	Panhandle consortium (ESU 13)	6B
6	Platteville, North Bend Central	6C
6	Medicine Valley	6D

Description of the models

On the following pages are concise descriptions of the model procedures. For more information about the details and results of the model process, the district named may be contacted or the ESU in which that district is located may be able to provide information about that model procedure.

Synopsis of Model Assessment Procedure for

Criterion 1

Model 1A

Criterion 1: Alignment with standards. Requires both alignment and sufficiency (that the assessment provides enough information to infer that the standard is met).

Synopsis of process for alignment:

This district used two methods to establish alignment of assessments with state standards. The two methods included a panel of district teachers and a panel of outside consultants that reviewed the assessments for alignment and confirmed the district's determination.

Synopsis of process for determining sufficiency:

Coverage was reviewed and determined to be sufficient by teaching staff trained for this purpose, and by outside experts.

Details of alignment process and sufficiency determination:

<u>Alignment</u>

This school district used panels of educators to a) write the standards and assessments, b) ensure that appropriate types of assessments were used, and c) that the content of the assessments matched state standards. Detailed information on the qualifications of the panel members and the results of the process was provided.

Subsequently, all teachers in grades 4, 8, and 11 reviewed the assessments during development for a match to the standards.

Two outside experts/consultants were engaged to review the content match, standards coverage, and the level of difficulty of the assessments. Reports from each consultant confirmed the alignment that was done by the district's teachers.

The district used special forms that were designed to guide test development and document the alignment of assessments and standards.

<u>Sufficiency</u>

The district documented the number and types of test items used in the district assessment. The outside experts also examined the assessment materials for alignment and confirmed that the assessments were sufficient to assess the standards.

Criterion 1

Model 1B

Criterion 1: Alignment with standards. Requires both alignment and sufficiency (that the assessment provides enough information to infer that the standard is met).

Synopsis of process for alignment:

Local standards were developed and were approved by the state as meeting or exceeding state standards. Local criterion-referenced assessments were developed for the local standards. Across several cycles of review and evaluation by teachers, the assessments were revised to assure alignment with the standards.

Synopsis of process for determining sufficiency:

Each assessment item and/or task was matched to the local standard it assessed, resulting in a table of the number of items/tasks for each standard. The number of score points for each standard (either single items for multiple choice, or rubric for constructed response) was determined to be adequate.

Details of alignment process and sufficiency determination:

Alignment

When a district has designed its own standards (approved as meeting or exceeding state standards), concerns about the alignment of the assessments with the standards remains a concern. Developing local standards almost certainly assures that the district's *curriculum* is aligned to its standards. This district used the same committee that developed the local standards to develop the assessments, helping to assure that the *assessments* would also be aligned to the standards. The committee was trained in assessment development with attention to alignment issues. At various times during assessment development, the committee sought assistance from outside assessment experts.

Once completed, the assessments were presented to all staff in the district for feedback and suggested revisions. Particular attention was paid to comments from teachers of grade levels one year below and one year above the assessed grades. Based on the teacher reviews and feedback, the committee revised the test items/tasks and the scoring guides/rubrics.

After the assessments were piloted and scored, teachers were again asked to comment on the alignment of the assessments to the local standards. The assessment committee reconvened to review and revise the items/tasks and the scoring rubrics in light of the pilot data and the teacher feedback.

Sufficiency

Each test item/task was mapped to the standard it assessed. A table was developed to show the total number of points for each standard. (The table also shows the number of points the district has decided are necessary for a student to have mastered the standard.) The number of points for each standard – one point for each multiple choice item and the number of rubric points for each short answer, extended response, or essay task – was then compared the "five point minimum" criterion.

Synopsis of Model Assessment Procedure for

Criterion 1

Model 1C

Criterion 1: Alignment with standards. Requires both alignment and sufficiency (that the assessment provides enough information to infer that the standard is met).

Synopsis of process for alignment:

A representative panel of qualified teachers from within the district judged the assessments to be matched to the standards and adequate to cover the standards. Meetings with teachers from other districts, organized through their ESU, gave district teachers a chance to share suggestions for assessments or parts of assessments and develop consensus on their understanding of the standards themselves.

Synopsis of process for determining sufficiency:

Items and tasks from the set of assessments were coded according to the appropriate standard, and a count was made. The criterion of at least 5 items per standard was used for objectively scored items. The criterion of at least two tasks per standard was used for subjectively scored assessments.

Details of alignment process and sufficiency determination:

Alignment

After developing the assessments, a representative panel of qualified teachers from within the district judged the assessments to be matched to the standards and adequate to cover the standards. The panel members all received assessment training prior to beginning the development process. A two-step process was used. First, grade level teachers wrote or selected assessments, using match to standards and coverage as one of the criteria. Next, teachers from other grade levels (who were also on the panel) reviewed these assessments for match to standards and adequacy of coverage, thus assuring that the reviews were conducted by teachers other than those who wrote or selected the assessments. A checklist was used for the review and was included in the portfolio.

Meetings with teachers from other districts, organized through their ESU, gave district teachers a chance to share suggestions for assessments or parts of assessments and develop consensus on their understanding of the standards themselves. While informal networking is not sufficient in itself to satisfy the criterion of alignment, the district found it helpful as an addition to their own formal development and review process.

Sufficiency

Items and tasks from the set of assessments were coded according to the appropriate standard and a count was made. The criterion of at least 5 items per standard was used for objectively scored items. The criterion of at least two tasks per standard was used for subjectively scored assessments. For subjectively scored assessments, scoring guidelines or rubrics were checked to insure that descriptions were clear.

Synopsis of Model Assessment Procedure for

Criterion 1

Model 1D

Criterion 1: Alignment with standards. Requires both alignment and sufficiency (that the assessment provides enough information to infer that the standard is met).

Synopsis of process for alignment:

This school district's alignment for grade 4 used a diverse teacher panel with an average of 10 years experience. The panel was involved in the development of the curriculum, the alignment of curriculum to standards, and the alignment of the assessments to standards. The principal and superintendent reviewed all alignment work.

Synopsis of process for determining sufficiency:

A diverse teacher panel determined: (a) which items measured which standards and (b) how many items per standard were needed to show adequacy of content/standards coverage.

Details of alignment process and sufficiency determination:

Alignment

For the Grade 4 standards a panel of local teachers took several steps to determine alignment of their Teacher Developed Tasks (e.g., Unit tests from SRA, ESU Reading Assessment System). These tasks included a) reviewing of 'best practices' research literature, existing curriculum, and current resources to identify four of the most important concepts taught in each subarea of Reading/Language Arts, (b) charting the concepts within subareas to determine the content progression over grades K-12, and (c) deciding which assessments measured which standards and filling in gaps with additional locally developed assessments. The panel worked in small and large group settings to reach consensus in each of the above 3 steps. All of the alignment work of the teacher panel was reviewed by the principal and superintendent.

Alignment of the District's CRT (Fluency Assessment) was provided in a report based upon a study conducted by the district. In the study, a panel of teachers from grades 4, 8, and high school provided ratings of the items in terms of high, moderate, and low levels of alignment of content to the standards. The teachers providing the ratings received training from outside consultants.

Sufficiency

To illustrate sufficiency of content coverage a chart was constructed that contained three column headings (Standards, Assessment Type, and Number of Items). The rows of the chart were the separate standards. The panel worked with each of the assessment types (e.g., NRT, CRT, unit tests, and teacher-created assessments) and listed the number of items from each assessment that measured a given standard.

Synopsis of Model Assessment Procedure for

Criterion 2

Model 2A

Criterion 2: Students have an opportunity to learn the content.

Synopsis of process for opportunity to learn:

This school district utilized multiple methods to determine whether its students had an opportunity to learn (OTL) the district assessment material. Methods included utilizing a computer software package, convening a panel of teachers that examined curriculum and lesson plans, and conducting classroom observations. The district realized and seized the opportunity to establish a strong staff development component as a part of its OTL analysis.

Details of process used to determine opportunity to learn:

A language arts committee examined local curriculum to ensure an alignment with the NE LEARNS. This information was entered into "Curriculum Designer," a software package which is accessible to teachers on the web. This gives teachers access to the curriculum, links to classroom activities, and web sites to assist teachers with the instruction of objectives that are linked to the standards.

This district conducted a series of workshops addressing relevant instruction aligned to state standards. Subsequently five departmental/grade level meetings were held to discuss a) the instructional strategies necessary to address standards, b) when standards would be taught during the school year, and c) to identify appropriate student work samples as evidence of progress toward standards.

All grade level language arts teachers were required to include information on standards on weekly lesson plans. In addition, administrators conducted random classroom observations to ensure instruction on standards.

Student indicator cards, listing information about individual student progress toward the state standards, were developed to provide a self-monitoring tool for teachers.

Synopsis of Model Assessment Procedure for

Criterion 2

Model 2B

Criterion 2: Students have an opportunity to learn the content.

Synopsis of process for opportunity to learn:

The district conducted activities in several ways to meet this criterion, and directly involved the judgment of both teachers and administrators. These activities included a curriculum review, an examination of sample lesson plans, and informal observations of classrooms by the principal.

Details of process used to determine opportunity to learn:

After the assessment was developed, many of the district's teachers were brought together as a curriculum committee to review opportunity to learn issues. This included having the teachers review the scope and sequence of the district's curriculum, and compare it to the coverage of the assessments.

The district systematically collected a sample of actual lesson plans at various times during the school year and had the curriculum committee review the lesson plans to see the extent to which students had been exposed to the material on the assessments prior to their administration.

During the year, the principal conducted "walk-arounds", noting whether or not instruction in the classrooms was linked to assessed topics. The principal also regularly reviewed lesson plans for evidence that students had the opportunity to learn the assessed outcomes. Where discrepancies were found between the principal's observations (either in the walk-around or from the lesson plans) and the assessments, feedback was provided to the teachers for adjustment and improvement.

Synopsis of Model Assessment Procedure for

Criterion 2

Model 2C

Criterion 2: Students have an opportunity to learn the content.

Synopsis of process for opportunity to learn:

A two-pronged approach was used: teacher surveys and a panel review of lesson plans. Noteworthy in this model is the survey of all appropriate teachers (not just those in grades 4, 8, and 11) and the use of a random selection (by standard) of just a few lesson plans to obtain verification of the survey's results without an enormous burden.

Details of process used to determine opportunity to learn:

Teacher surveys were prepared, listing the standard number and skills taught for grades 4, 8, and 11, respectively. For each survey, space was provided for listing the dates that content was taught, according to lesson plans. All fourth grade teachers and a random sample of third grade teachers responded to the 4th grade survey. All 7th and 8th grade Language Arts and Reading teachers and 8th grade Social Studies teachers responded to the 8th grade survey. All members of the high school English department responded to the 11th grade survey. The rationale for selection of teachers to be surveyed was based on identifying teachers who, according to the district curriculum, taught content or skills relevant to the standards. All teachers who were surveyed responded.

Survey findings reported for each grade level indicated that 100% of the relevant content had been taught prior to assessment. Noteworthy was the report of the fourth grade teachers who found that although their students had all been taught the appropriate content and skills, the amount of time given to different content varied depending on the teacher. This led to discussions about instructional methods within the grade level that resulted in at least two benefits: a) professional development activity that was targeted to specific, identified needs and b) 4th grade teachers reported feeling less isolated from one another.

The second method used was to verify survey results with a "reality check" of lesson plans to see whether what was reported was indeed taught. In order to do this without placing an enormous burden on teachers, a small sample of lesson plans was selected. The same teachers who had received the survey were randomly assigned one of the standards (from those they taught) and were asked to turn in a lesson plan for panel review. Three different panels (3 people each for 4th, 8th, and 11th grade standards) reviewed these plans to verify that they did match with the particular standard and that the lessons occurred prior to assessment.

Synopsis of Model Assessment Procedure for

Criterion 2

Model 2D

Criterion 2: Students have an opportunity to learn the content.

Synopsis of process for opportunity to learn:

Several sources were used to determine whether district students had an opportunity to learn the material to be assessed. These sources included curriculum guides, teacher-principal conferences, lesson plans, texts and other instructional materials. A panel of representative teachers made many of these judgments.

Details of process used to determine opportunity to learn:

This district used three strategies to judge the extent that students had the opportunity to learn the content of the standards prior to the assessment. One strategy involved a diverse team of teachers representing the grade levels in question (with an average of 10 years teaching experience). This team came together to a) examine the curriculum guides for alignment with the curriculum and to determine when the material was taught and b) examine and tag textbooks and other instructional materials used to develop classroom assessments in terms of where and when the standard was to be assessed.

A second strategy included highlighting lesson plans to show what standard was being assessed by a given lesson. The lesson plans were examined by the principal to determine if the lessons were taught prior to the assessment.

The third strategy included teacher-principal conferences held to determine what standards were taught and when by each teacher.

In the portfolio, the district provided illustrations of sample lesson plans and highlighted where and when the standard was taught within the plan in the areas of Spelling, Reading, and English for grades 3-6.

Synopsis of Model Assessment Procedure for

Criterion 3

Model 3A

Criterion 3: The assessments are free from bias or offensive situations

Synopsis of process for bias/sensitivity review:

This district relied heavily on intensive training of a multiracial/ethnic staff to understand and detect bias in assessment materials.

Details of bias/sensitivity review process:

The district identified and selected a panel that was reflective of the race/ethnicity of the district.

Panel members received intensive orientation and training in understanding and detecting bias related to gender, ethnic, cultural, socioeconomic, and religious factors in assessment materials.

The process included an in-depth rating of all test items, including extensive reviewer comments on every test item, which were then used to revise or delete items.

A wide variety of training materials was used and reflected in the district's assessment portfolio.

Synopsis of Model Assessment Procedure for

Criterion 3

Model 3B

Criterion 3: The assessments are free from bias or offensive situations

Synopsis of process for bias/sensitivity review:

This district trained assessment writers in awareness of item bias issues, convened a panel to review early drafts of the assessment items and tasks, and conducted a statistical analysis of the results of the assessment for evidence of biased items.

Details of bias/sensitivity review process:

All assessment writers were required to participate in assessment training, which included attention to minimizing test and item bias. The bias training was comprehensive and included a review of the definitions of test and item bias, examples of biased assessments, and guided practice to recognize and correct bias.

Additionally, a panel of teachers and community members was convened to review the assessments for potential bias. This panel reflected much of the diversity that is present in the student and parent communities. Namely, the panel was balanced by gender, was diverse by ethnicity, and had members from the community's various religious groups. Each panel member, after receiving the same bias training that was provided to the teacher assessment writers, independently evaluated every assessment item and task for bias. An "index", defined as 20% of the panel judging an item/task as biased, was used to flag items/tasks for review.

Finally, after the assessments were administered and scored, statistical analyses were conducted to search for evidence of bias. Inclusion of any of these analyses would add to the quality of a district's bias review. Differential Item Functioning (DIF) analysis was conducted for the multiple-choice items. A DIF analysis compares the differential performance of students in two groups (for example: male vs. female, white vs. non-white, no free/reduced lunch vs. free/reduced lunch) on an item-by-item basis, and provides a statistical test of the magnitude of the bias, if any. The items flagged by DIF as possibly biased were then compared to the panel's judgment. One item emerged as needing revision, and will be re-written before the next administration of the assessment.

Synopsis of Model Assessment Procedure for

Criterion 3

Model 3C

Criterion 3: The assessments are free from bias or offensive situations

Synopsis of process for bias/sensitivity review:

A panel of teachers, administrators, paraprofessionals, parents, and other community members reviewed the district-wide criterion referenced assessments for evidence of bias or offensive language. Revisions were made based on the findings. The test was edited so that names, places, occupations, and the like represented the culture of the local community.

Details of bias/sensitivity review process:

A panel of teachers, administrators, paraprofessionals, parents, and other community members was formed, representing the ethnic composition of the community. The distribution of the panel, by category (teacher, parent, etc.) and ethnicity was reported in the district's assessment portfolio. This panel reviewed the district-wide criterion referenced assessments for evidence of bias or offensive language. This bias review focused on both gender equity and cultural sensitivity. The gender review involved coding each test question according to whether the pronouns and other references were male or female. The cultural review involved identifying names, places, occupations, celebrations and customs and checking for their relevance to the student population taking the exam.

Revisions were made based on the panel's findings. Test items were edited to achieve an equitable distribution of "male" and "female" items. The test was edited so that names, places, occupations, and the like represented the culture of the local community and did not reinforce any ethnic or cultural stereotypes.

Synopsis of Model Assessment Procedure for

Criterion 3

Model 3D

Criterion 3: The assessments are free from bias or offensive situations

Synopsis of process for bias/sensitivity review:

This district used a panel composed of members from various local community groups (e.g., ACLU, Domestic Rights) and members with varying socioeconomic and cultural perspectives to evaluate the items. A workshop on how to evaluate items for bias was held for item writers and the bias review panel. Panel members used forms to comment on possible item bias.

Details of bias/sensitivity review process:

A set of materials describing aspects of potential item bias was used to train item writers and a bias review panel. These materials were an adaptation from copyrighted materials produced by National Evaluation Systems, Inc.

A diverse panel of community leaders conducted the bias review using the set of materials described above. Letters from an external review committee were presented in the district's assessment portfolio indicating which items had problems and why. Comments about individual items detailed the nature of the problem and the steps that were taken to remedy it.

Synopsis of Model Assessment Procedure for

Criterion 4

Model 4A

Criterion 4: The level is appropriate for students

Synopsis of process for level of assessment:

This model relied heavily on ESU expertise to train a cadre of veteran teachers who then a) examined the developmental and cognitive appropriateness of assessment materials and b) conducted readability analyses using the Readability Master 2000 software package.

Details of process used to determine the level of the assessment(s):

Veteran panels of teachers representing grades 4, 8, and 11 received training in examining the developmental and cognitive appropriateness of test materials. They then reviewed and selected appropriate texts and items for district assessments.

Forty additional teachers reviewed the assessments to confirm the appropriateness of the assessments for students in grades 4, 8, and 11.

Subsequently, the same veteran panels of teachers were instructed in the use an application of readability formulas. Specifically, the Readability Master 2000 software package was used, which includes the Dale-Chall Readability Formula, the Spache Readability Formula, and the Fry Graph for Estimating Readability. The panels used the software package to analyze and modify test materials that did not meet the criterion of developmental appropriateness.

Synopsis of Model Assessment Procedure for

Criterion 4

Model 4B

Criterion 4: The level is appropriate for students

Synopsis of process for level of assessment:

This district conducted readability analyses of the assessments, judged the adequacy of the readability indices, and implemented a procedure to adjust the readability of assessments that were found to need adjustment.

Details of process used to determine the level of the assessment(s):

Before the assessments were finalized, appropriate readability indices were selected for each grade level assessed. Text passages were analyzed using the indices and the results were reported in terms of the average readability and the range of readability indices. After readability indices were applied, the results were evaluated, and in some cases assessment items/tasks were adjusted when an index formula suggested the item/task was inappropriate.

The district recognized that different levels of readability would be acceptable for a reading passage than for a set of instructions in an assessment.

In some cases the readability was found to be farther off grade level than was judged to be appropriate by the district's staff. When this occurred, the text was re-written and re-analyzed to ensure that it was age appropriate.

Synopsis of Model Assessment Procedure for

Criterion 4

Model 4C

Criterion 4: The level is appropriate for students

Synopsis of process for level of assessment:

Representative panels of qualified teachers and other educators within the district, including curriculum specialists, judged the assessments from a developmental and cognitive perspective. This occurred in three waves, with different panels reviewing the assessments, thus allowing "different pairs of eyes" to examine developmental appropriateness.

Details of process used to determine the level of the assessment(s):

The first level of review was completed by a district curriculum committee composed of representatives from each grade level. The composition of this committee (content, grade, years of experience, years of experience in the appropriate grade) and the nature of the assessment training it received were described in the assessment portfolio. The first layer of review was built into the assessment development and selection process in that developmental appropriateness formed one of the criteria used. This review encompassed all assessments to be used in reporting assessment results to the state. The second level of review occurred in departmental or grade level meetings, as part of regular staff development time. The review process employed a consensus methodology. The composition of each of the relevant departmental (English/Language Arts) or grade level meetings' participants was described in the assessment portfolio. School psychologists reviewed the work of each grade level or department panel, providing another review external to the committee consensus process. The third level of review involved a repeat review by the district's curriculum committee. At this stage, an attempt was made at final resolution of all conflicts and questions. This was achieved for all levels except one grade level, where additional review and possible adjustments for reading level were suggested for consideration in the following year.

Synopsis of Model Assessment Procedure for

Criterion 4

Model 4D

Criterion 4: The level is appropriate for students

Synopsis of process for level of assessment:

This district used two different groups to study the reading/writing assessments for appropriate level and conducted a readability analysis.

Details of process used to determine the level of the assessment(s):

Two different groups were used to study the reading/writing assessments to determine if they were appropriate for each grade level: a panel of educators (e.g., teachers representing K-12 in four groupings, special education and Title I reading teacher, and an administrator) and workshop trainers. The panel of educators and the workshop trainers used an assessment rubric (an Assessment Critique Sheet[1]) to analyze and determine the appropriate grade level of the items. The analysis of the items from the trainers was returned to the panel of educators for resolution or revision. Three different readability analyses were used to estimate the reading level of the assessments: Spache (K-3), Dale-Chall (4-8), and Fry (9-12). When the readability level for an item or passage was found to be too high for a given grade level, the text was modified. A computer printout of the results was included in the portfolio. The printout gave not only the grade level for a passage, but also listed unfamiliar words for a given grade level. [1]The Assessment Critique Sheet contained 10 questions rated on a 4-point scale adapted from SBE Design Team (12/99) to rate each assessment.

Synopsis of Model Assessment Procedure for

Criterion 5

Model 5A

Criterion 5: There is consistency in scoring. There are two parts for this, one for objectively scored assessments and a second for subjectively scored assessments.

Synopsis of process for scoring consistency for objectively scored assessments:

This district uses only objectively scored assessments to meet the state standards. Therefore it only used an internal consistency reliability procedure to estimate consistency in scoring.

Synopsis of process for scoring consistency for subjectively scored assessments:

N/A

Details of process used to determine the level of the assessment(s)

Objectively scored assessments This district created its assessments with "Classroom Manager," a software test item data base package marketed by CTB/McGraw-Hill. The software package reports estimates of internal consistency reliability (e.g., Cronbach's Alpha, KR20) for assessments it produces. The district reported the statistics from their analyses of the assessments and reported how it will use this information to improve their assessments. Subjectively scored assessments: N/A

Synopsis of Model Assessment Procedure for

Criterion 5

Model 5B

Criterion 5: There is consistency in scoring. There are two parts for this, one for objectively scored assessments and a second for subjectively scored assessments.

Synopsis of process for scoring consistency for objectively scored assessments:

The district collected and reported multiple estimates of score consistency, including internal consistency, test-retest estimates, and the correlation coefficient between the district's assessment and a norm-referenced test. These various estimates collectively supported the notion that scores on the assessment are reliable.

Synopsis of process for scoring consistency for subjectively scored assessments:

Scoring rubrics were designed by a panel of experienced district staff and were piloted by district teachers before being incorporated into curriculum guides. Before administration of the assessments, teachers were trained in scoring using "anchor" papers to an acceptable level of consistent scoring.

Details of process used to determine the level of the assessment(s):

<u>Objectively scored assessments</u>

The district routinely uses a software package (SPSS) to calculate an estimate of internal consistency (Cronbach's Alpha) on multiple-choice assessments. The district also routinely calculates estimates of test-retest reliability (coefficients of correlation for scores on separate testings of the same students on the same test) and reports a range of these reliability estimates for varying lengths of time between testing occasions.

An additional estimate of the reliability of an assessment can be obtained from the correlation between the assessment and some other measure of similar learning outcomes. The district calculated and reported correlations between the district's assessment and the norm-referenced test used at each of the three grade levels. This correlation provides an indication of the extent that students would be classified in the same way on the two tests.

<u>Subjectively scored assessments:</u>

The district developed scoring rubrics for its classroom-based assessments using a panel of teachers and curriculum specialists from within the district. The development process included checking the clarity and adequacy of the scoring rubrics through pilot testing the assessments. Teachers involved in the pilot test suggested revisions of the rubrics before they were published in the district's curriculum "frameworks" documents.

Once a part of the district's "frameworks" documents, the assessment rubrics were incorporated into district-wide professional development activities for language arts team leaders and department heads; follow-up training and monitoring activities for other certified staff are carried out at the building level. The district also reported the level of inter-rater agreement for the subjectively scored assessments they use.

Synopsis of Model Assessment Procedure for

Criterion 5

Model 5C

Criterion 5: There is consistency in scoring. There are two parts for this, one for objectively scored assessments and a second for subjectively scored assessments.

Synopsis of process for scoring consistency for objectively scored assessments:

Reliability was estimated as the consistency of decisions about student performance levels on two different assessments measuring the same standard.

Synopsis of process for scoring consistency for subjectively scored assessments:

Reliability was estimated as the consistency of decisions about student performance levels on two different assessments measuring the same standard.

Details of process used to determine the level of the assessment(s):

<u>Objectively scored assessments</u> For the standards assessed by the norm-referenced test used by the district, there were two sources of information about each student on the same standard (the NRT and another assessment, either the CRA or a classroom assessment). In this small district, classroom teachers compared the achievement category (standard met or not met) of each student on both assessments to ascertain whether there was a match in classification. The percent of students who were classified the same way by both assessments (where both assessments indicated the student met the standard, or both assessments indicated that the student did not meet the standard) was used as a measure of scoring and decision consistency. <u>Subjectively scored assessments</u>: The same strategy employed for objectively scored assessments was also employed for subjectively scored assessments.

Synopsis of Model Assessment Procedure for

Criterion 5

Model 5D

Criterion 5: There is consistency in scoring. There are two parts for this, one for objectively scored assessments and a second for subjectively scored assessments.

Synopsis of process for scoring consistency for objectively scored assessments:

A consortium of districts participated in training on item writing, item analyses, and item critiquing from external consultants to enhance score consistency prior to administration. The consortium used Cronbach's Alpha an estimate of internal consistency.

Synopsis of process for scoring consistency for subjectively scored assessments:

A panel of teachers from the consortium received training on performance assessments and rubric development. Once assessments and rubrics were developed a different group of teachers reviewed them. Holistic scoring was used to score the assessments and a sample of assessments was double scored to estimate the level of inter-rater agreement.

Details of process used to determine the level of the assessment(s):

Objectively scored assessments

1. Scores from a representative sample of students from the consortium was selected to compute estimates of reliability (consistency) using internal consistency methods. The results presented in the assessment portfolio varied from quite low to high across the assessments. The district will be making revisions to the assessments with coefficients of .50 or less (one district chose .70 or less as their criterion). The district also included the number of items for each reliability estimate to provide additional information about the reliability estimates. This allowed reviewers to understand why some reliability estimates may have been higher or lower.
2. Because the NRT and CRT assessments were matched to the standards correlation coefficients were computed between the scores of students on these two assessments suggested consistency of agreement between two different assessments. Moderate to high correlations indicated the same students were similarly classified (indicating consistency of classification). A spreadsheet was used to perform these calculations.

Subjectively scored assessments

Only performance assessments that were given in at least 8 districts in the consortium, administered in same manner, and had at least 50 student performances available for analysis were used to determine inter-rater scoring consistency.

1. Panel of teachers representing all grade levels, special education and Title I, with over six years experience, received assessment training. The training included developing performance assessments and scoring guides (rubrics). Draft assessments and rubrics were reviewed by teachers who were not on the panel and suggested revisions as needed. After administration of the assessments, notes were made of relative strengths and weaknesses of the assessments.
2. Holistically scored ratings of a sample of students' performance found exact inter-rater agreement of 94% and adjacent agreement of 100% using a 4-point scale.

Synopsis of Model Assessment Procedure for

Criterion 6

Model 6A

Criterion 6: Mastery levels are appropriate

Synopsis of process for setting mastery levels:

This district used a panel of educators to examine test items and estimate the level of performance that a barely master student would obtain on its assessments. The method is called a Modified Angoff Method.

Details of process for setting mastery levels:

The district formed a panel of teachers that was balanced for gender and educational background. The panel received training in the Angoff standard setting method from a principal who had previously been trained at the ESU in the method. The Angoff method is a judgmental approach to standard setting that requires panels to examine each test item individually and estimate how a student on the borderline between performance levels will perform on that item. Item performance estimates are averaged across the teachers and then summed over all the items to obtain the mastery level for a performance category.

The district's assessment portfolio included the forms and training materials used by the district and the results of the process.

Synopsis of Model Assessment Procedure for

Criterion 6

Model 6B

Criterion 6: Mastery levels are appropriate

Synopsis of process for setting mastery levels:

A consortium of small districts identified the curricular areas that were common across districts and aggregated student data across the districts for those areas. A student-based standard setting procedure was applied to these across-district data. Individual districts then used these cut scores as a basis for locally-determined adjustments to the levels of performance required for mastery.

Details of process for setting mastery levels:

Several districts, representing small districts within a large geographic area, collaborated to establish cut scores (mastery levels) for their assessments. Setting cut scores can be problematic when the number of students tested is small. This may require a similar collaboration if a district uses a data based method.

The districts agreed on the assessments that were commonly occurring across the districts. Results from those assessments were aggregated into a single set of data.

Cut-scores were established using multiple strategies. The "Professional Estimates" method relied on the judgments of the teachers of the students from the individual districts and the "Borderline Group" method served as a check of the "Professional Estimates" method. These multiple strategies supported each other to better inform the final cut score decision.

After cut-scores were set, individual districts were encouraged to use the cut-scores as guidelines for establishing locally defined performance levels. This maintains consistency across the districts to a large extent, while allowing individual districts to adjust the scores to match particular ways in which the district's curriculum is unique.

Synopsis of Model Assessment Procedure for

Criterion 6

Model 6C

Criterion 6: Mastery levels are appropriate

Synopsis of process for setting mastery levels:

This district was part of an ESU-sponsored standard setting study conducted by an external consultant. The consultant used the "Borderline Group" method, a student-based method that will assist in the mastery level decision-making process. Each member of the consortium applied the results in his or her own district.

Details of process for setting mastery levels:

The ESU sponsored a mastery level study for a consortium of regional schools. An external consultant was hired to perform the study. The consultant used the borderline group method. The borderline group method is a student-based method. Judgments about which students "just" fall into each performance level are collected, then the average (or median) score for those groups of students are used to establish the scores associated with each level of performance.

After students took the tests, teachers provided data on their judgments of student proficiency level, and the consultant used those ratings and actual student performance on the assessments in a "Borderline Group" study to set standards for selected assessments at grades 4, 8, and 12. In this way, the consultant was able to use a larger sample of students for the study than any one district could have provided alone. The consultant's report formed the basis for setting mastery levels in the district.

The district used the mastery levels set in the external consultant's study for those assessments that had been studied. For assessments that had not been used in the external study, the district did a smaller scale borderline group study, using district teachers and data, to establish cut scores for each performance level used within the district. It used those results as the mastery levels for those assessments, realizing that adjustments might need to be made as more data becomes available in future years.

Synopsis of Model Assessment Procedure for

Criterion 6

Model 6D

Criterion 6: Mastery levels are appropriate

Synopsis of process for setting mastery levels:

A student-focused approach was used to set mastery levels using teacher judgment of student proficiency for each assessment. Two student classifications were used: masters and nonmasters for reporting to the state and four performance levels (beginning, progressing, proficient, and advanced) were used for Title I reporting.

Details of process for setting mastery levels:

A team of teachers projected how their students might perform on each of the assessments by identifying students on the borderline between performance categories. The average score for these students was used to set the mastery level or score for the assessment. The same procedure was used to set the mastery level for the other performance levels (beginning, progressing, proficient, and advanced) if there was sufficient students classified as borderline between each of the performance categories.